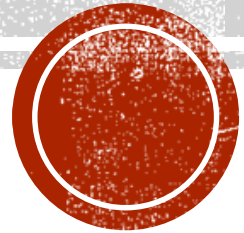
CALIFORNIA HOUSE PREDICTION

Created by : Dimas Sigit Priyatna



AGENDA

1. Context
2. Yang ditawarkan
3. Kenapa Machine Learning ?
4. Modelling (Algoritma)
5. Kesimpulan
6. Rekomendasi

1. CONTEXT

What ? **Data Science Baru**	**Why ?** **Memprediksi harga rumah**
When ? Saat ini	Where ? Perumahan di California
Who ? Profesional Data Science	How ? Machine LEarning

2. YANG DITAWARKAN

MODEL MACHINE LEARNING

CLOSE THE DEAL

3. KENAPA MACHINE LEARNING ??

Untuk meminimalisir error ketika hendak membeli rumah di California

DATASET

```
df = pd.read_csv('data_california_house.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14448 entries, 0 to 14447
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   longitude           14448 non-null  float64
 1   latitude            14448 non-null  float64
 2   housing_median_age  14448 non-null  float64
 3   total_rooms         14448 non-null  float64
 4   total_bedrooms      14311 non-null  float64
 5   population          14448 non-null  float64
 6   households          14448 non-null  float64
 7   median_income       14448 non-null  float64
 8   ocean_proximity     14448 non-null  object 
 9   median_house_value  14448 non-null  float64
dtypes: float64(9), object(1)
memory usage: 1.1+ MB
```

longitude	--> Garis bujur, semakin tinggi semakin mejauh dari barat
latitude	--> Garis yang horizontal / mendatar semakin tinggi semakin menjauh dari utara
housing_median_age	--> Rata-rata usia rumah, semakin besar maka semakin tua
total_rooms	--> Total ruangan
total_bedrooms	--> Total kamar
population	--> Total populasi
households	--> Total kepala keluarga
median_income	--> Rata-rata pemasukan perkepala keluarga (US$)
median_houseValue	--> Rata-rata harga rumah (US$)
ocean_proximity	--> Jarak ke pantai/laut

4. MODELING(ALGORITMA)

Evaluasi Performa Tiap Algoritma

	Model	Mean_RMSE	Std_RMSE	Mean_MAE	Std_MAE	Mean_MAPE	Std_MAPE
0	Linear Regression	-57632.542826	1849.848109	-40067.601495	879.827346	-0.229984	0.007046
1	KNN Regressor	-52524.475864	877.181946	-35241.754781	640.979322	-0.194576	0.002963
2	DecisionTree Regressor	-65459.023136	3122.190567	-43994.461868	1523.861907	-0.252010	0.008113
3	RandomForest Regressor	-46987.033129	1781.410678	-30936.557489	804.529409	-0.171841	0.005370
4	XGBoost Regressor	-44777.380269	1657.923858	-29327.414366	895.569645	-0.163405	0.002532

MENCARI BENCHMARK

- RMSE

	Model	CV_RMSE	Mean_RMSE	Std_RMSE
0	Linear Regression	[54502.06262, 56556.79807, 59358.32257, 59094.05374, 58851.47713]	57632.54283	1849.84811
1	KNN Regressor	[50900.85796, 53260.36138, 53084.44285, 52299.57833, 53077.1388]	52524.47586	877.18195
2	DecisionTree Regressor	[66493.50262, 61541.0298, 69822.49535, 62250.79337, 67187.29455]	65459.02314	3122.19057
3	RandomForest Regressor	[44166.52126, 45848.41261, 48958.10211, 47367.03507, 48595.0946]	46987.03313	1781.41068
4	XGBoost Regressor	[42104.87555, 43759.2023, 46755.52234, 45329.39245, 45937.90871]	44777.38027	1657.92386

Berdasarkan nilai RMSE, jika dilihat dari nilai rata-rata nya, XGBoost memilki nilai yang paling baik, kemudian diikuti dengan RandomForest. namun jika dilihat dari standard deviasi, KNN regressor merupakan yang paling baik, diikuti dengan Linear Regression

- MAE

	Model	CV_MAE	Mean_MAE	Std_MAE
0	Linear Regression	[38583.04338, 39649.41065, 40350.5091, 40654.24576, 41100.79858]	40067.80150	879.82735
1	KNN Regressor	[34461.31122, 38152.27931, 35357.59312, 34583.78639, 35653.80387]	35241.75478	640.97932
2	DecisionTree Regressor	[44167.03297, 42310.27527, 46138.13187, 42273.18681, 45083.68242]	43994.46187	1523.86191
3	RandomForest Regressor	[29714.89589, 30776.40988, 31380.02172, 30670.43161, 32141.02834]	30936.55749	804.52941
4	XGBoost Regressor	[28068.27861, 29225.49421, 30207.20045, 28894.95893, 30441.14163]	29327.41437	895.56965

> Berdasarkan nilai MAE, jika dilihat dari nilai rata-rata nya, XGBoost memilki nilai yang paling baik, kemudian diikuti dengan RandomForest. namun jika dilihat dari standard deviasi, KNN regressor merupakan yang paling baik, diikuti dengan RandomForest

- MAPE

	Model	CV_MAPE	Mean_MAPE	Std_MAPE
0	Linear Regression	[0.22641, 0.22521, 0.22166, 0.23979, 0.23686]	0.22998	0.00705
1	KNN Regressor	[0.19251, 0.19769, 0.19033, 0.19434, 0.19801]	0.19458	0.00296
2	DecisionTree Regressor	[0.26106, 0.24182, 0.25421, 0.24317, 0.25978]	0.25201	0.00811
3	RandomForest Regressor	[0.17107, 0.16967, 0.16562, 0.17103, 0.18182]	0.17184	0.00537
4	XGBoost Regressor	[0.16153, 0.163, 0.16277, 0.16142, 0.16831]	0.16341	0.00253

> Berdasarkan nilai MAPE, jika dilihat dari nilai rata-rata nya, XGBoost memilki nilai yang paling baik, kemudian diikuti dengan RandomForest. Jika dilihat dari standard deviasi XGboost juga merupakan yang paling baik, diikuti dengan KNN regressor
>
> Selanjutnya akan dilakukan prediksi pada test set dengan menggunakan 2 model benchmark terbaik yaitu XGboost dan RandomForest

PREDICT TEST SET (TUNED MODEL)

Melakukan prediksi pada test set dengan menggunakan model XGBoost dan hyperparameter terpilih.

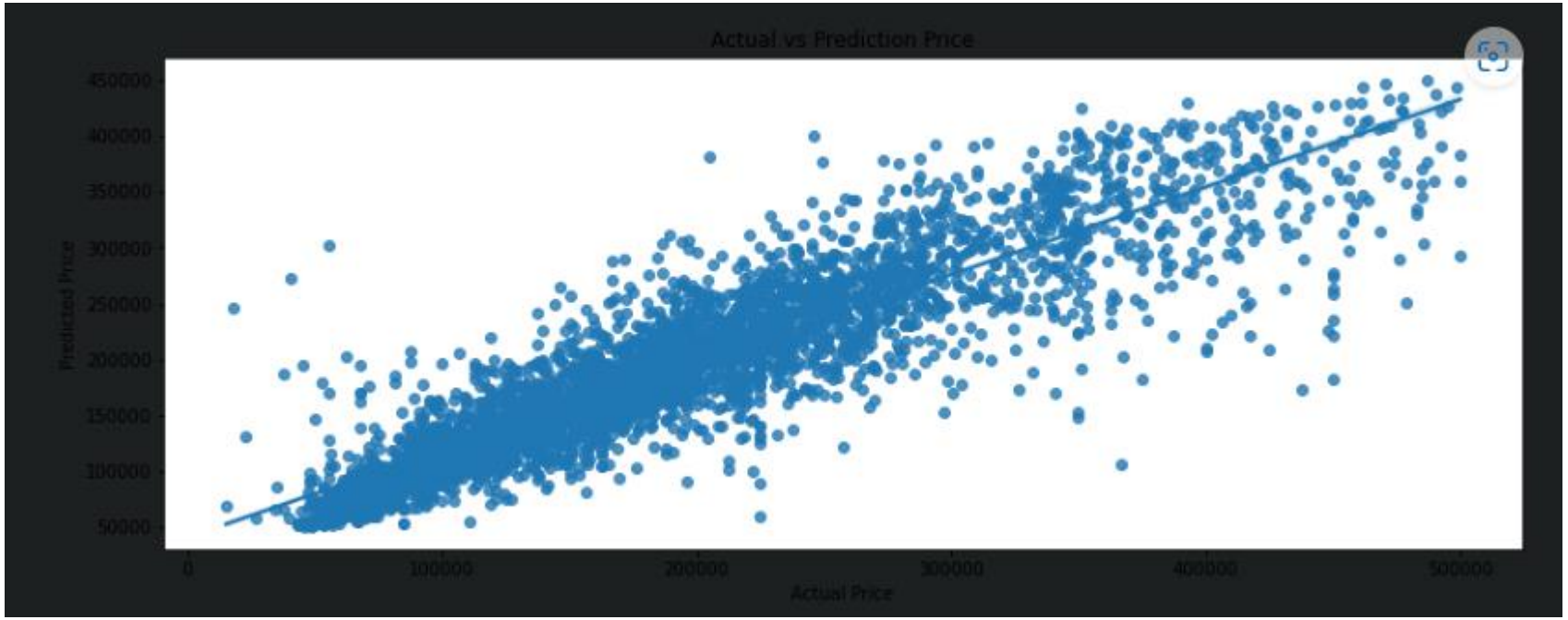
	RMSE	MAE	MAPE
XGB	42005.532508	28387.586877	0.168875

PERFORMA

Model mengalami peningkatan performa (nilai RMSE, MAE & MAPE berkurang) dengan dilakukannya hyperparameter tuning.

- **Sebelum Tuning** :
  - RMSE : 43951.740457
  - MAE : 29369.021115
  - MAPE : 0.167447
- **Setelah Tuning** :
  - RMSE : 43634.316318
  - MAE : 29092.170909
  - MAPE : 0.164377

PERTANDINGAN HARGA YANG DIPREDIKSI VS AKTUAL



FEATURES IMPORTANCES

5. CONCLUSION

Berdasarkan pemodelan yang sudah dilakukan, fitur 'ocean_proximity' dan 'median_income' menjadi fitur yang paling berpengaruh terhadap 'median_house_value'.

Hal ini cukup wajar artinya kita dapat mengkonfirmasi bahwa lokasi ternyata masih menjadi predictor yang paling kuat dalam menentukan harga suatu rumah. semakin rumah tersebut berada dalam area / kawasan yang elit, tentu saja harga rumah nya akan tinggi dan juga sebaliknya. Dalam kasus ini rumah yg berada di kawasan pinggir dengan view laut merupakan rumah yang paling mahal dibandingkan dengan rumah yang berada di lokasi lainnya.

Hal ini juga berbanding lurus dengan fitur median income, dimana rata-rata penghasilan seseorang dalam suatu area akan menentukan harga rumah di sekitarnya. Semakin besar rata-rata penghasilan seseorang di area tersebut, maka akan semakin mahal harga rumahnya, begitu pula sebaliknya.

Jika kita melihat berdasarkan nilai RMSE, didapati nilai RMSE cukup tinggi, hal ini dikarenakan metric RMSE memiliki beberapa kelemahan: RMSE tergantung oleh scala dari data, jadi semakin besar skala, maka nilai RMSE nya juga besar. RMSE juga dipengaruhi oleh outlier, semakin banyak outlier maka RMSE juga bisa semakin besar. Seperti yang kita ketahui data kita memiliki outlier yg cukup banyak, tapi jika outlier nya dihilangkan maka kita akan loss informasi yang banyak pula. Oleh karena itu pada kasus ini saya lebih melihat hasil pemodelan menggunakan metric MAPE yang tidak terlalu sensitive terhadap adanya outlier, dimana hasil dari metric MAPE sendiri yg sebesar 16% yang artinya persen kesalahan hasil prediksi data dibanding data actual hanya sekitar 16%. Selain itu nilai MAPE 16% artinya termasuk kedalam kategory 'Good Forecast' atau model peramalan baik.

6. RECOMENDATIONS

Hal-hal yang dapat dilakukan untuk mengembangkan model agar lebih baik lagi :

1. Penambahan fitur-fitur yang memiliki korelasi langsung dengan harga suatu rumah, misal luas rumah, fasilitas rumah, perusahaan developernya , dll.

2. Data perlu diperbaharui karena data yang digunakan merupakan data yang sudah lama yaitu tahun 1990. data ini tentu saja sudah sangat tidak relevan dengan kondisi pada saat ini. karena adanya faktor inflasi dan sebagainya.

3. Dari sisi modeling mungkin dapat ditingkatkan dengan metode hyperparameter yang lebih baik seperti gridsearch. Metode gridsearch mencoba seluruh kombinasi hyperparameter. Sedangkan pada randomized search yang kita gunakan dalam model tidak semua kombinasi hyperparameter dicoba tetapi kita memilih secara acak dari seluruh kemungkinan kombinasi.

4. Model ini dapat digunakan untuk prediksi harga perumahan yang memiliki fitur sejenis dengan dataset California house. Karena jika dilihat dari perbandingan nilai train dan test nya, performa model cukup stabil artinya model cenderung tidak overfitting/underfitting. Namun perlu diingat kembali bahwa data ini merupakan harga rumah di tahun 1990, yang tentu saja akan jauh berbeda dengan harga rumah di tahun sekarang, ini berkaitan dengan range harga harga yang akan diprediksi, karena jika range nya melewati atau diluar range harga dalam model, maka hasilnya akan menjadi bias.